Data Management Plans for Regional Associations

1 Introduction

This discussion is offered to motivate and guide preparation of a Regional Association Data Management Plan (RA-DMP) by each of the IOOS Regional Associations (RAs). The IOOS DMAC will regularly update this white paper based on plans submitted by the RAs and results from IOOS Regional Development Workshops.

RA-DMPs will contain: parameter-level descriptions of data and model output collected and served within the region, assessments of the region's system capabilities relative to the key IOOS DMAC Plan elements (access, transport, online browse etc.), documentation of standards and practices employed, assessments of how well the data management system meets the goals of the Regional and national IOOS, a listing of identified gaps, a prioritized list of remedies, and a statement of future plans.

Each RA should follow a similar outline when preparing their plans to facilitate comparison of common elements between regions. The following sections discuss: the expected benefits of a RA-DMP, the role of the related RA Conceptual Design/System Design document for establishing priorities, guidance for performing assessments, defining gaps, and for the expression near-term and long-term plans. This paper concludes with suggested elements of the RA-DMP and a template for contents in an appendix.

2 Purpose for an RA-DMP

The RA-DMPs, and their subsequent updates[1], will serve several purposes for planning and implementation of the broader IOOS enterprise[2]. They are:

- Describe the current status of the implementation of IOOS-compatible data management systems within each Regional Association,
- Describe near-term plans[3] for implementing additional or expanded capabilities,
- Identify gaps in data management capabilities that will require resources or assistance to resolve.

[1] "subsequent updates" implies these plans are updated on some regular cycle. An annual cycle makes sense from an Ocean.US perspective, but the workshop may usefully discuss this. What cycle can be useful without inducing planning fatigue?

[2] "IOOS enterprise" is the author's short-hand to indicate all IOOS participating elements.

[3] The wording "near-term plans" is deliberately undefined. The time horizon selection should consider the multi-year funding profile available -- among other factors. Items included in near-term plans would be ones the RAs believe they can accomplish. Periods of 18-36 months may be appropriate, but may differ for each RA. The significance is to clearly distinguish between planned actions and gaps.

It may be useful to think of these plans as a way for the RAs to communicate a self-assessment of their current data management status, near-term plans, and gaps. With these plans in place, it will become possible to:

- Identify opportunities for collaborative development of system elements,
- Identify opportunities for reuse of successful implementations by others,
- Prioritize work on enterprise-wide solutions where similar gaps exist, and
- Produce fact-based assessments of enterprise-wide progress on implementation of the Data Management & Communication component of IOOS.

Obviously there are other mechanisms to share this type of information – workshops, websites, newsletters, and participation in joint projects. These types of efforts will continue and are important. The formality of these periodic plans, however, ensures there will be a basic level of shared knowledge among the IOOS participants.

Creating these plans also is a process intended to assist the RAs with management of their local enterprise. Assessing the present status is basically a fact-gathering exercise (although it can be a challenge to organize those facts into an understandable presentation). But, selecting near-term plans and prioritizing gaps requires judgment based on local needs and allocation of available resources. *Consequently, it is important that these Data Management Plans involve whatever decision-making structure the RA has adopted for prioritizing and allocating its resources.* Publication then provides a basis for common understanding of the RA plans and priorities among the RA participants with respect to data management.

Why use an RA-level focus? Such planning could be done at various organizational levels within the IOOS enterprise. For example, some RAs operate as a 'federation' of sub-regional organizations. It is logical to document similar plans at the sub-regional level. However, this approach suggests using plans at the RA level as the best way to aggregate the information for use by the broader IOOS enterprise. Individual RAs might ask their sub-regional elements to prepare similar planning materials, but that is left to local decisions.

The intent of these plans is to assess past, current and future progress of development with respect to IOOS DMAC guidance and standards as applied to RA, national IOOS and global GOOS needs. Plans should be expressed as a fairly high-level summary of the end-to-end implementation of their data management systems. Inventories of observed parameters may require a finer granularity. However, plans need not include details such as operating procedures, reference guides or code. Such information may be included or referenced in the Plans if the RAs wish. Expanded coverage on specific implementations may be included as an appendix.

3 Data Management Priorities

A Conceptual Design/System Design Plan for RA/RCOOS is a companion document to the RA-DMP. The Conceptual-System Design Plan will likely be prepared by each RA's management body with input from their associated councils and committees. The content of Conceptual Design/System Design plan has a critical role to the RA-DMP as it defines the broader purpose and business design of the RCOOS. Data management will be a key element to reaching the broader system objectives. However, it is only a necessary (but not sufficient) component. The challenge for the data manager is to design, develop and implement capabilities that serve those objectives -- both at the regional level and the IOOS enterprise level. The oft-expressed frustration is the lack of clearly stated system objectives (requirements), and their prioritization, at both the regional and national levels.

Development of these Data Management Plans, especially in terms of deciding what to do first, should reflect the priorities and objectives of the regional-level system design work. Given that this regionalized design effort is not complete, the remainder of this discussion focuses on a generalized approach to creating these plans for those elements that all RAs have in common. Making them specific to the objectives of each RA will be a task for participants returning to their regions after the workshop.

With this obvious mandate to address regional priorities, data managers should also consider priorities as expressed by the IOOS DMAC Steering Team process. Some of these will be reflected in the current version of the "Guide for IOOS Data Providers"[4]. The Guide will be updated as enterprise-level planning proceeds (which specifically includes development of these regional data management plans). Standards and best practices identified in the Guide as either 'proposed', 'recommended' or 'development work' should be given appropriate consideration for implementation or development work at the regional level – especially where it aligns with the regional priorities. This discussion will later make some suggestions for priorities based on the current version of the Guide, but regional planning should stay alert for updates.

An example may help clarify this use of the Guide. Suppose the Guide identifies a data transport and access capability needing additional work, (e.g. TAET 6.0 "Help develop conventions for serving unstructured grids"). If an RA has a local need for this capability, then the RA might choose to identify development of those conventions as part of its near-term plans. Successful development would lead to a 'proposed' and eventually 'recommended' standard for that capability. When multiple RAs or other IOOS participants target such an item for development, then Ocean.US can act to ensure collaboration among them.

Another source of information on data management priorities will be federal funding opportunity announcements or BAAs issued by IOOS-participating agencies. The DMAC planning at Ocean.US will be working with agencies to include high priority tasking in their future announcements, some of which may come from the 'gaps'

[4] Available at http://dmac.ocean.us/dacsc/guidance02.jsp

identified by these regional data management plans and similar planning by other IOOS enterprise participants.

This process implicitly reflects an evolutionary development and implementation of the IOOS data management component – some might call it a 'distributed spiral development'. An alternative process would be a more centralized top-down development that offers less risk for successful integration across the enterprise. Some capabilities may be developed with the centralized model and provided to participants as services they can access or software for deployment at local installations. The choice of which path to take will be made on a case-by-case basis.

4 Assessing Present Status

There is no single cookbook for conducting an assessment. But the following steps should be considered. Some of the details may not be relevant to the local region's situation. In the end, however, the planner should have a comprehensive list of data flows produced or consumed for regional needs, and have an understanding of the 'as-is' data management capability for each.

4.1 Gathering the information

Begin by identifying each regional provider of data, model output, or analysis product. Gather information on each stream, cite existing documentation and confirm as necessary with appropriate personnel. (Many regions have begun this process for near real-time observations through their participation in the IOOS Regional Observation Registry Project). Suggested types of information include the following, much of which would be available from existing metadata descriptions if they have been written:

- **Parameters/variables** – noting when these align with the 20 'IOOS Core Variables' and the additional list of Atmospheric and Terrestrial Observations in the IOOS Development Plan, Part II, page 18. A similar crosscheck to the variable lists for coastal GOOS would be helpful[5]

- **Format/Content** – what format or content standard or convention applies to this data flow? Look for file format descriptions if held in a flat-file repository, or a data dictionary if held in a database structure.

- **Sources** – observing systems, types of platforms, or analysis product

- **Temporal and spatial coverage** –
  - Begin and end dates ("present" if continuing)
  - Geographic extent

[5] Tables II.3a, II.3b, II.7 in Annex II of the Implementation Strategy for the Coastal Module of GOOS

    - Horizontal resolution or distribution
    - Vertical resolution or distribution
    - Temporal characteristics, such as sampling frequency

- **Data Volumes** – quantitatively estimate size of present holdings and growth rate (both as stations or records and as volume in 'bytes').

- **Data Latency** – especially for continuing data flows, characterize the time lag between observation and availability at the IOOS access point.

- **Versioning** – are there multiple versions of the same data? The DMAC Plan anticipates three generic versions – real-time, delayed-mode, and archive.

- **QA/QC –** is quality control applied, where is its description, what is the data flagging convention(s)?

- **Vocabulary –** what vocabulary standards or conventions are used for variable names and units?

- **Access Methods –** how are these data made available to the IOOS enterprise?

- **Access/Use Restrictions** – are there any restrictions on full and open access? For example, is a holding period or permissions required?

- **Metadata** – if metadata records have been created, where are they posted and are they searchable?

- **Archive** – how has archiving of these data been addressed?

- **Developmental Status** – Research, Pilot, Pre-Operational, or Operational as defined in the IOOS Development Plan, Part I, page 8. This can be relevant because some enterprise-level data management requirement vary with this status.

- **Future Plans** – Does the data flow 'owner' have plans to modify their services regardless of any IOOS activity? It is important that the regional IOOS data management function be are aware of any such plans.

4.2 Assessing the Capabilities

Three types of assessment are needed. The first judges if the observations, model outputs, and products available to the region are adequate to support the regional and national and global goals for observing systems. The second assesses whether each of the required elements for machine-to-machine interoperable systems, as identified in the IOOS DMAC Plan, have been addressed and implemented. The third assesses whether these implementations are compliant with IOOS DMAC standards and best practices.

Obviously the second and third assessments are related; the former is a check that a capability exists, the latter is a check that it is implemented in an compliant way. They will be discussed together.

Depending on how your RA is organized, the first assessment – that of adequacy of the system to meet regional, national and global needs, may fall outside the data management unit, (e.g. too few observations of a variable). However, the information collected in section 4.1, would form the basis of such an assessment and the data management group is probably best positioned to assemble the list. Adequacy in this context can be difficult to judge. Ideally, the system requirements at both the regional and enterprise levels would be specific enough to make this a simple pass/fail testing process. A more nuanced grading scheme is needed. It is suggested that the 'assessor' simply create a list of how the existing capability falls short. Admittedly that may be a largely subjective judgment. The list of shortfalls becomes the input to the step of prioritizing gaps and creating a near-term plans. Both are relevant to data management plans. The results of the assessment, regardless of who makes it, should appear in the RA-DMP.

The second assessment looks to see whether or not software systems have been deployed to implement the key data management functional capabilities as defined in the DMAC Plan[6] These capabilities are bundled into the following categories:

- Transport and Access
- Metadata
- Archive
- Data Discovery
- Uniform On-line Browse

Two additional categories have become increasingly significant as DMAC planning has evolved:

- IT security
- QA/QC

The third assessment checks to see that deployed systems adhere to standards. The 'Guide for IOOS Data Providers' contains current information on these standards. The following is some guidance on specific aspects to check based on the current version of the 'Guide'. Please refer to the 'Guide' itself rather than relying on this abbreviated presentation. Note that the current Guide also carries forward previously published 'Concrete Guidance' documents in its appendices.

It is worth remembering that the overarching vision in all these capabilities is to facilitate interoperability of data flows across the IOOS enterprise. The goal is to eliminate, or at least reduce, barriers to routine machine-to-machine exchange of the data. The need for 'manual format busting', 'web-scraping', 'manual search and select retrieval' or similar manpower intensive or potentially unstable process requirements should become obsolete

[6] See especially DMAC Plan, Part II, Section 1, beginning on page 65

as IOOS DMAC matures. Another general criterion might be to consider whether any User A would normally have to call or email Provider B in order to understand how to access and use the data. If so, then some aspect of the IOOS data management capability is not yet adequately implemented.

- **Transport and Access** - The current Guide identifies fourteen implementations of transport and access – of which half are categorized as developmental. First determine which are potentially relevant to the data flow being considered. For the relevant ones, assess the how well the existing data flow address the desired capability. Note what additional work that may be needed. The focus of the current Guide is on capabilities related to the DAP, OGC and OBIS (DiGIR server) protocols. Success with other techniques; such as Z39.50 servers, FTP sites, or direct SQL calls to remote RDBMS installations; should be noted.

- **Metadata –** the current Guide identifies eleven implementations of metadata – of which one is categorized as developmental work. Some are very clear, suitable for a pass/fail judgment -- such as MET 3.0 concerning use of a standard IOOS vocabulary in the keywords. Given the community's historical abhorrence of the metadata topic, it is expected that most 'data flows' will need additional work in this area.

- **Archive –** the current Guide carries forward four implementation of archive from previous IOOS guidance. Nothing new has been added to the discussion in the existing DMAC Plan. AET 1.0 and 4.0 in the Guide specifically apply to regional data center operations, so assessments in these areas are particularly relevant.

- **Data Discovery** – the current Guide is silent in this area. The DMAC Steering Team has suggested that this topic be consolidated into the purview of Metadata. There is logic in claiming that a proper Metadata Catalog and Search/Query services (as described in Sections 3.3 and 4.0 of the DMAC Plan, Part II) will satisfy this capability. The assessment here might consider how any existing capability serves these objectives. Evidence of even partially successful implementations would be valuable information.

- **Uniform On-Line Browse –** the current Guide is silent in this area. The DMAC Steering Team has postponed further discussion until a later time. The earlier discussion in the DMAC Plan identifies 'Live Access Servers' and GIS-enabled Web Mapping Services as likely prototypes for fulfilling this need. The description in the DMAC Plan (Part I, page 39) makes the following statement italicized below. Information on successful implementations would be valuable. <u>However, be careful not to confuse On-Line Browse capabilities with Data Transport and Access capabilities</u>.

  The DMAC Plan anticipates that many IOOS data providers will host metadata-enabled, open-source, or commercial on-line browse tools for end users. In

addition, effective management of the DMAC Subsystem requires a system-wide view of IOOS data — the ability to visualize and assess all IOOS data in a uniform manner. The Uniform On-Line Browse capability of DMAC must provide geo- and time-referenced graphics and data in human-readable tables. It will use the DMAC Data Transport services for access to IOOS data. It must be accessible through standard web browsers. The DMAC Subsystem must provide a seamless segue from Data Discovery to Uniform On-line Browse.

- **IT Security** – the current Guide contains four statements, all of which are carried forward from previous guidance. Only three are applicable to regional activities – SEC 1.0, 2.0 and 3.0. Comments on regional experience with these issues would be valuable. For additional perspective, review the discussions of IT Security in the DMAC Plan – Overview, Part I, page 45.

- **QA/QC –** the current Guide carries forward two statements from previous guidance. Both might be considered as 'punts'-- deflecting the issues to other groups or entities for specification. However, a word search for 'quality control' in the DMAC Plan yields some forty references – indicating it is acknowledged as a significant topic for IOOS. But as noted in the Executive Summary section of the DMAC Plan (page 5)

  "[…. There are system components that] transfer data from the measurement platforms to … the locations at which Primary Data Assembly and Quality Control (PDA&QC) occur. The systems that convey data from sensors to primary data centers/sites lie outside the scope of the current DMAC Plan."

In essence this implies that QA/QC is viewed to occur before the data reaches the DMAC component of the system.[7] But DMAC also acknowledges the requirement to handle data of "known quality", implying that results of any QA/QC will be kept with the data. In this perspective, we are asking regional planners to note whether QA/QC is available with each of their data flows.

5 Gaps and Near-Term Plans

The assessments in Section 4.1 should reveal additional items (data streams and products) needed to meet regional, national IOOS, and GOOS priorities. The assessments related to IOOS DMAC capabilities and adherence to standards should reveal gaps that can be addressed by the Region internally, and those which will require additional work by the community or by IOOS DMAC. These remedies should be prioritized and plans developed by the Regions using processes developed by themselves according to their organizational structure. These plans should take into consideration available and potential resources or referred to IOOS DMAC if local solutions do not seem possible.

The near-term plans should include items in progress or items that the RA are able to do but are waiting for resource allocation. The needed resources may be called a shortfall.

[7] This 'denial' of QA/QC responsibility by the DMAC component was motivated by a realization that the technical expertise of the Steering Team was inadequate to evaluate the adequacy of QA/QC requirements for the variety of observing systems within IOOS. It was hoped that this task could be taken on by an "OMAC"-type group ['O' indicating Observations] that never materialized.

A shortfall of this type could be time, money, staffing, etc. Gaps may include items in progress or items planned for future implementation but have shortfalls such as a standard that comes up slightly short to meet requested needs. The plan may also suggest ways in which gaps could be fixed or bridged.

6 Data Management Report Elements

The Data Management Plan should include the inventory lists and assessments mentioned in Section 4. The document should include gaps and prioritized lists of proposed solutions and near-term activities. Current and future operations details should be included in summary form. Details of staffing, space, hardware, storage, overhead may be included as this information may be useful to other regions who are in the process of establishing data management capabilities. RCOOS details beyond data management should appear in the associated Conceptual Design/System Design plan.

Details given above then should be presented along timeline (GANTT) charts. It is worth reminding the planner that attempting to plan in the future is very difficult. It is expected that RA Data Management plans are snapshots and subject to change, especially the future.

An example outline for the RA-DMP is supplied in Appendix A.

Appendix A - Suggested Data Management Plan Outline

The following is a suggested outline for a Regional Association Data Management Plan document. It parallels the outline used earlier by the Alaska Ocean Observing System, and the topical breakout of the IOOS DMAC Plan[8]. Authors should add new sections or sub-sections as necessary to further describe their regional situation. Appendices or references to other published documents are encouraged as a way to incorporate substantive details.

Also suggest that plans use a five-year perspective. The focus is on the near-term. What can be implemented in 12-24 months, while also identifying longer-term needs. Annual updates to the plan allow changes in direction or emphasis as changes in technology or resources may present new opportunities or new constraints.

AOOS Data Management Plan

Executive Summary

1. Introduction
2. Regional Inventory of data, models, and products
3. Metadata and Data Discovery
4. Transport and Access
5. Archive
6. IT Security
7. Quality Assurance/Quality Control
8. Gaps
9. Plans
10. Staff and facilities
11. GANTT Charts

References

Appendix A: Terms, definitions and acronyms

Appendix B: QARTOD QA/QC flags for data sets

[8] Data Management and Communication Plan for Research and Operational Integrated Ocean Observing System, Ocean.US Publication No.6, 2005